BITS Pilani
Pilani | Dubai | Goa | Hyderabad
Work Integrated Learning Programmes

# Assignment: Predicting Attrition

294 - VIMAL KUMAR N, KRITIKA SHARMA, LAL CHAND SHARMA

Introduction to Data Science

M.Tech Data Science and Engineering

## Overview

- **Objective**: Analyzing and Building models for Predicting Attrition
- **Methodology**:
  - EDA: Perform exploratory analysis of the data
  - Preprocessing the collected data: Perform data wrangling / Pre-Processing to improve outcomes
  - Analyzing the dataset: The most important features that push an employee to leave the organization are detected
  - Balancing the dataset: Since the dataset is not already balanced, it is necessary to be equalized
  - Building the predictive model: The suitable configuration for the model is selected to increase the prediction accuracy (Logistic regression and Decision tree to predict)
  - Validating the model: Compare the performance of the two classifiers – Logistic regression and Decision tree to predict

## Methodology

- EDA: Perform exploratory analysis of the data
- Preprocessing the collected data: Perform data wrangling / Pre-Processing to improve outcomes
- Analyzing the dataset: The most important features that push an employee to leave the organization are detected
- Balancing the dataset: Since the dataset is not already balanced, it is necessary to be equalized
- Building the predictive model: The suitable configuration for the model is selected to increase the prediction accuracy (Logistic regression and Decision tree to predict)
  - Logistic regression:
    - For better performance we used **feature scaling** in Logistic regression
    - Applying **Recursive Feature Elimination (RFE)** for feature selection in Logistic regression
  - Decision Tree:
    - Performed Class Imbalance Check
    - Applied UNDERSAMPLING
    - Applied OVERSAMPLING
- Validating the model: Compare the performance of the two classifiers – Logistic regression and Decision tree to predict

## Results

- Table for the evaluation metric for each ML technique used
- Logical regression:

| Accuracy | 86.5 | | | | |
|---|---|---|---|---|---|
| | | precision | recall | f1-score | support |
| | 0 | 0.865248 | 0.987854 | 0.922495 | 247 |
| | 1 | 0.75 | 0.191489 | 0.305085 | 47 |
| accuracy | | | | 0.860544 | 294 |
| Macro avg | | 0.807624 | 0.589672 | 0.61379 | 294 |
| Weighted avg | | 0.846824 | 0.860544 | 0.823794 | 294 |

BITS Pilani
Pilani | Dubai | Goa | Hyderabad
Work Integrated Learning Programmes

# Assignment: Predicting Attrition

294 - VIMAL KUMAR N, KRITIKA SHARMA, LAL CHAND SHARMA

Introduction to Data Science

M.Tech Data Science and Engineering

## Dataset

- Data columns (total 33 columns):
  - dtypes: float64(2), int64(17), object(14)
  - memory usage: 379.1+ KB
  - There are 1,470 rows and 33 columns in the data.
- Single file used (Final dataset Attrition-1.csv)
- ['Date_of_termination', 'Unnamed: 32'] columns have only one Unique Values and drop them
- ['BusinessTravel', 'NumCompaniesWorked', 'StockOptionLevel','TrainingTimesLastYear'] columns are removed due to usability
- There are 1 column to remove due to high correlations (['MonthlyIncome']
- Balanced or imbalanced – what is the distribution
- 20 % data is distributed as Training set, 80% used as testing set

## Results

- Recursive Feature Elimination (RFE) for feature selection

| Accuracy | 86.5 | | | | |
|---|---|---|---|---|---|
| | | precision | recall | f1-score | support |
| | 0 | 0.867857 | 0.983806 | 0.922201 | 247 |
| | 1 | 0.714286 | 0.212766 | 0.327869 | 47 |
| accuracy | | | | 0.860544 | 294 |
| Macro avg | | 0.791071 | 0.598286 | 0.625035 | 294 |
| Weighted avg | | 0.843307 | 0.860544 | 0.827189 | 294 |

- Decision Tree:

| Accuracy= 76.87 | | | | |
|---|---|---|---|---|
| Classification report- | | | | |
| | precision | recall | f1-score | support |
| 0 | 0.87 | 0.85 | 0.86 | 247 |
| 1 | 0.30 | 0.34 | 0.32 | 47 |
| accuracy | | | 0.77 | 294 |
| macro avg | 0.59 | 0.60 | 0.59 | 294 |
| weighted avg | 0.78 | 0.77 | 0.77 | 294 |

  - Method 1: UNDERSAMPLING

| Accuracy= 55.79 | | | | |
|---|---|---|---|---|
| Classification report- | | | | |
| | precision | recall | f1-score | support |
| 0 | 0.57 | 0.54 | 0.55 | 48 |
| 1 | 0.55 | 0.57 | 0.56 | 47 |
| accuracy | | | 0.56 | 95 |
| macro avg | 0.56 | 0.56 | 0.56 | 95 |
| weighted avg | 0.56 | 0.56 | 0.56 | 95 |

# Assignment: Predicting Attrition

294 - VIMAL KUMAR N, KRITIKA SHARMA, LAL CHAND SHARMA

Introduction to Data Science

M.Tech Data Science and Engineering

## Feature Engineering Techniques

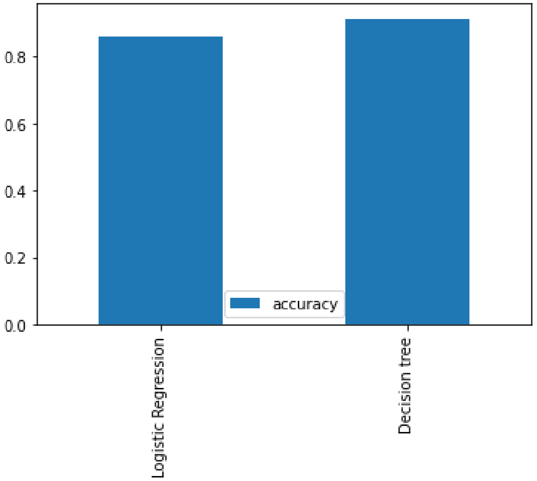
- ['Date_of_termination', 'Unnamed: 32'] columns have only one Unique Values and drop them
- ['BusinessTravel', 'NumCompaniesWorked', 'StockOptionLevel','TrainingTimesLastYear'] columns are removed due to usability
- There are 1 column to remove due to high correlations (['MonthlyIncome']
- Implemented feature encoding using label encoder for below features:
- ['Attrition', 'Department', 'Gender', 'JobRole', 'MaritalStatus', 'OverTime', 'Higher_Education', 'Date_of_Hire', 'Status_of_leaving', 'Mode_of_work', 'Work_accident', 'Source_of_Hire', 'Job_mode']
- ['Attrition'] is selected as feature

## Results

○ Method 2: OVERSAMPLING

| | | | | |
|---|---|---|---|---|
| Accuracy= 91.3 | | | | |
| Classification report- | | | | |
| | precision | recall | f1-score | support |
| 0 | 0.97 | 0.85 | 0.91 | 247 |
| 1 | 0.87 | 0.98 | 0.92 | 247 |
| accuracy | | | 0.91 | 494 |
| macro avg | 0.92 | 0.91 | 0.91 | 494 |
| weighted avg | 0.92 | 0.91 | 0.91 | 494 |

- Plot of the curves



- Conclusion

We can see that Decision tree has ~6% better accuracy than Logistic regression but Decision tree is an Oversampling model hence we will select Logistic regression.